Using the New SURVEY Procedures From a Modeling Perspective

Jonas V. Bilenas, Banking Industry SME

PhilaSUG Fall 2019 Meeting
October 22, 2019

This presentation is an update from (Bilenas 2009).

Outline

- Many new SURVEY procedures were added in SAS8® and SAS9® and many updates have been made in recent SAS/STAT versions.
- A Few Applications:
  - Generating Samples:
    - Simple samples
    - Stratified samples
  - Regression Model Validations:
    - Bootstrapping
    - K-Fold Validations
  - SURVEY Modeling Procedures to Get Correct p-values for Sampled Data. We will focus on PROC SURVEYREG in this presentation.
  - An Occasional Trivia Question.

Note: All code was generated and tested on SAS Studio SAS® OnDemand for Academics. Most of the code presented in this presentation is influenced from David Cassell through his SAS-L comments and his SAS User Group Presentations on SURVEY PROCS (2006, 2007).

Trivia Question

SAS® has been generating many new versions of Base SAS. Other SAS packages have version numbers that don't necessarily align with BASE SAS version.

- Which of these can be used and gives you the most information? Note that option (**d**) is a valid SAS procedure but the full code is not listed in the answer below.

    a) Look at the LOG

    b) PROC SETINIT; run;

    c) PROC PRODUCT_STATUS; run;

    d) PROC EXPLODE; ~;

Trivia #1

- Scores

a)	Look at the LOG	2 **points**
b)	PROC SETINIT; run;	**3 points**
c)	PROC PRODUCT_STATUS; run;	**6 points**
d)	PROC EXPLODE;	**1 point**

SETINIT: Answer (b): 3points. LOG:

Product expiration dates:

---Base SAS Software	**31DEC2020 (CPU A)**
---SAS/STAT	**31DEC2020 (CPU A)**
---SAS/GRAPH	**31DEC2020 (CPU A)**
---SAS/ETS	**31DEC2020 (CPU A)**
---SAS/OR	**31DEC2020 (CPU A)**
---SAS/IML	**31DEC2020 (CPU A)**
---SAS/QC	**31DEC2020 (CPU A)**
---SAS/CONNECT	**31DEC2020 (CPU A)**
---SAS Enterprise Miner	**31DEC2020 (CPU A)**
---SAS Integration Technologies	**31DEC2020 (CPU A)**
---SAS/Secure 168-bit	**31DEC2020 (CPU A)**
---SAS Enterprise Miner Server	**31DEC2020 (CPU A)**
---SAS Enterprise Miner Client	**31DEC2020 (CPU A)**
---SAS Credit Scoring	**31DEC2020 (CPU A)**
---SAS Text Miner	**31DEC2020 (CPU A)**
---SAS High-Performance Forecasting	**31DEC2020 (CPU A)**
---SAS Enterprise Guide	**31DEC2020 (CPU A)**
---Forecast Server Conditional Setinit	**31DEC2020 (CPU A)**
---OR OPT	**31DEC2020 (CPU A)**
---OR PRS	**31DEC2020 (CPU A)**
---OR IVS	**31DEC2020 (CPU A)**
---OR LSO	**31DEC2020 (CPU A)**
---SAS/ACCESS Interface to PC Files	**31DEC2020 (CPU A)**
---SAS/ACCESS Interface to MySQL	**31DEC2020 (CPU A)**
---SAS Forecast Studio	**31DEC2020 (CPU A)**
---SAS Forecast Server Mid-Tier	**31DEC2020 (CPU A)**
---SAS/IML Studio	**31DEC2020 (CPU A)**
---SAS Workspace Server for Local Access	**31DEC2020 (CPU A)**
---SAS Workspace Server for Enterprise Access	**31DEC2020 (CPU A)**
---SAS/ACCESS to Postgres	**31DEC2020 (CPU A)**
---High Performance Suite	**31DEC2020 (CPU A)**
---SAS Add-in for Microsoft Excel	**31DEC2020 (CPU A)**
---SAS Time Series Workspace Macros	**31DEC2020 (CPU A)**

PRODCT_STATUS: 6 points. LOG Output:

```
71 options nocenter fullstimer;
72 proc product_status;
73 run;
For Base SAS Software ...
Custom version information: 9.4_M6
Image version information: 9.04.01M6P110718
For SAS/STAT ...
Custom version information: 15.1
For SAS/GRAPH ...
Custom version information: 9.4_M6
For SAS/ETS ...
Custom version information: 15.1
For SAS/OR ...
Custom version information: 15.1
Image version information: 9.04.01M6P050819
For SAS/IML ...
Custom version information: 15.1
For SAS/QC ...
Custom version information: 15.1
For SAS/CONNECT ...
Custom version information: 9.4_M6
For SAS Enterprise Miner ...
Custom version information: 15.1
For SAS Time Series Workspace Macros ...
Custom version information: 15.1
Image version information: 9.04.01M5P110718
For SAS/ACCESS to Postgres ...
Custom version information: 9.4_M6
For SAS Integration Technologies ...
Custom version information: 9.4_M6
For SAS/Secure 168-bit ...
Custom version information: 9.41_M3
```

```
For SAS Credit Scoring ...
Custom version information: 15.1
For SAS Text Miner ...
Custom version information: 15.1
For SAS High-Performance Forecasting ...
Custom version information: 15.1
For High Performance Suite ...
Custom version information: 2.2_M7
For SAS Forecast Server Mid-Tier ...
Custom version information: 15.1
Image version information: 9.04.01M5P110718
For SAS/ACCESS Interface to PC Files ...
Custom version information: 9.4_M6
For SAS/ACCESS Interface to MySQL ...
Custom version information: 9.4_M6
```

<u>Trivia #1 (d) PROC EXPLODE</u>
<u>1 point for answer (d).</u>

```
*   *    *    ****   ****   *   *         *   *    *    *      *       ***   *   *  *****  *****  *   *
*   *   * *   *   *  *   *  *   *         *   *   * *   *      *      *   *  *   *  *      *      **  *
*   *  *   *  *   *  *   *   * *          *   *  *   *  *      *      *   *  *   *  *      *      * * *
*****  *****  ****   ****     *           *****  *****  *      *      *   *  *   *  ****   ****   *  **
*   *  *   *  *      *        *           *   *  *   *  *      *      *   *  * * *  *      *      *   *
*   *  *   *  *      *        *           *   *  *   *  *      *      *   *  ** **  *      *      *   *
*   *  *   *  *      *        *           *   *  *   *  *****  *****   ***   *   *  *****  *****  *   *


****    ***    ***     *      *      *
 *  *  *   *  *   *    *      *      *
 *  *  *   *  *   *    *      *      *
 ***   *   *  *   *    *      *      *
 *  *  *   *  *   *    *      *      *
 *  *  *   *  *   *
****    ***    ***     *      *      *
```

NOTE: PROC EXPLODE is not ODS Compliant.

/* http://support.sas.com/documentation/onlinedoc/base/91/explode.pdf */

Sampling Data using PROC SURVEYSELECT

DATA WE WILL USE FOR PRESENTATION

DATA OPTIONS FOR BANKING AND FINANCE FOLKS

- TOY DATA
- SIMULATED DATA
- DATA NOT NECESSARILY RELATED TO BANKING AND FINANCE

Did Someone Say Sampling?

Did Someone Say Sampling?

- **Data was taken from Michael Jackson's Complete Guide to Single Malt Scotch, 4th Edition. Running Press.**
- **<u>No alcohol will be served during this presentation.</u>**

Where is Whiskey or Whisky Made?

Lots of countries are producing Whiskey or Whisky including the following as top producers:

- **Germany**
- **Taiwan**
- **Finland**
- **Australia**
- **India**
- **Canada**
- **Japan**
- **Ireland**
- **USA**
- **Scotland**

- **SOURCE:**
  - **https://usaspiritsratings.com/en/blog/insights-1/top-whiskey-producing-countries-of-the-world-99.htm**

Data Used in This Presentation:

- **Variables:**
  - Name of Single Malt Scotch: **WHISKY**
  - **REGION**
  - **Age**? What is it referring too? Missing values were removed but may be included in a future presentation from the 5th edition from Michael Jackson.
  - **Alcohol** by Volume
  - Special **Wood**?
  - **Rating** (100 point scale) provided by Michael Jackson

Take a Sample from the Text

```
proc freq data=scotch.scotch ; /* Not Yet Sampled */
  tables region/missing;
run;
```

region				
region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cambeltown	6	1.64	6	1.64
Highlands	292	79.78	298	81.42
Islay	42	11.48	340	92.90
Lowlands	26	7.10	366	100.00

Generate Random Samples

- You want to feature 10 bottles to sample in a party or in a marketing test.
- Use SURVEYSELECT Procedure

```
options nocenter fullstimer;

libname SCOTCH '~';

proc surveyselect
     data=scotch.scotch
     method=srs
     /*rate=.10*/
     N=10
     out=sample1 seed=201910;
run;
```

For this presentation we will mainly focus on these 2 sampling options:

- **METHOD=SRS**
  - **Simple Random Sampling without replacement**
- **METHOD=URS**
  - **Unrestricted random sampling with replacement**
- **STAT 15.1 has many new methods which are listed in appendix A.**

Appendix A: STAT 15.1 Additional METHOD OPTIONS FOR SURVEYSELECT

- BALBOOTSTRAP
- BERNOULLI
- POISSON
- PPS
- PPS_BREWER
- PPS_MURTHY
- PPS_SAMPFORD
- PPS_SEQ

- Others in the support.sas.com:

https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_surveyselect_syntax01.htm&docsetVersion=15.1&locale=en#statug.surveyselect.selectmethod

Generate Random Samples

Output

Selection Method	Simple Random Sampling

Input Data Set	SCOTCH
Random Number Seed	201910
Sample Size	10
Selection Probability	0.027322
Sampling Weight	36.6
Output Data Set	SAMPLE1

Generate Random Samples

```
proc print data=sample1 noobs; run;
```

whisky	region	age	alcohol	wood	rating
CRAIGELLACHIE	Highlands	15	62.7		76
DEANSTON	Highlands	16	55		71
GLENDRONACH	Highlands	18	43		78
GLENFARCLAS	Highlands	12	43		87
GLENGRANT	Highlands	5	40		65
GLENMORANGIE	Highlands	21	46		85
JURA	Highlands	20	54		78
MACALLAN	Highlands	15	43	Sherry	92
MILTONDUFF	Highlands	10	40		75
MORTLACH	Highlands	22	65.3		85

Next Trivia Question:

- According to Rick Wicklin (2013), in a DATA STEP and PROC IML, which function should **now** be used :

  a) RANUNI, RANNOR, RANBIN, and other "RANXXX" functions

  b) RAND function along with the CALL STREAMINIT(*seed*)

Trivia Answer: (b) for 5 points

Reference is from his blog site (The DO LOOP):

"Six reasons you should stop using the RANUNI function to generate random numbers"
https://blogs.sas.com/content/iml/2013/07/10/stop-using-ranuni.html

- **Example of RAND and STREAMINIT are illustrated in (Bilenas and Tahiliani, 2016 SESUG).**

Take a Stratified Sample: 2 from each Region

```
proc sort data=scotch.scotch out=region
          noequals tagsort force;
  by region;
run;

proc surveyselect data=region
                  method=srs
                  n=2
                  out=sample2
                  seed=1874;
  strata region;
run;

proc print data=sample2 noobs;
run;
```

Next Trivia Question:

- What does the PROC PRINT option **NOOBS** do?

  a) Do not print any observations.

  b) Do not drink any booze while coding.

  c) Do not print the observation number in the output.

Trivia Answer:

- What does the PROC PRINT option **NOOBS** do?

  Answer is **C**: Do not print the observation number in the output.

  5 points.

Take a Stratified Sample

Selection Method	Simple Random Sampling
Strata Variable	region

Input Data Set	REGION
Random Number Seed	1874
Stratum Sample Size	2
Number of Strata	4
Total Sample Size	8
Output Data Set	SAMPLE2

region	whisky	age	alcohol	wood	rating	SelectionProb	SamplingWeight
Cambeltown	SPRINGBANK	25	46		95	0.33333	3
Cambeltown	SPRINGBANK	32	59.54		94	0.33333	3
Highlands	TULLIBARDINE	27	45		77	0.00685	146
Highlands	GLENLOCHY	32	47.9		69	0.00685	146
Islay	ARDBEG	10	40	Sherry	85	0.04762	21
Islay	LAPHROAIG	10	57.3		88	0.04762	21
Lowlands	AUCHENTOSHAN	22	43		86	0.07692	13
Lowlands	LITTLEMILL	25	53.5		76	0.07692	13

Can we predict the Rating with a Regression Model?

Some Exploratory Analysis (FREQS)

region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Cambeltown	6	1.64	6	1.64
Highlands	292	79.78	298	81.42
Islay	42	11.48	340	92.90
Lowlands	26	7.10	366	100.00

wood				
wood	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	336	91.80	336	91.80
B/S	1	0.27	337	92.08
Bourbon	2	0.55	339	92.62
Oak	1	0.27	340	92.90
Port	1	0.27	341	93.17
Rum	1	0.27	342	93.44
Sherry	24	6.56	366	100.00

```
Title FREQS on Full CLASSES of Region and WOOD;
PROC FREQ data=scotch.scotch;
  tables region wood/missing;
run;
title;
```

Some Exploratory Analysis (FREQS)

The FREQ Procedure

wood	wood Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing	336	91.80	336	91.80
OTHER WOOD	6	1.64	342	93.44
SHERRY	24	6.56	366	100.00

Collapsing the number of Wood categories

The FREQ Procedure

wood	wood Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing	336	91.80	336	91.80
Specifed	30	8.20	366	100.00

	N	NMiss	Min	P5	P25	P50	P75	P95	Max	Mean	Std
age	366	0	3	8	12	16	21	30	50	17.43	7.07
alcohol	366	0	40	40	40	43	54	61	65	46.82	7.59
rating	366	0	57	70	76	79	84	91	96	79.55	6.30

Some Exploratory Analysis (FREQS)

```
proc format;
  value $wood
  ' '      = 'Missing'
  'Sherry' = 'SHERRY'
   other   = 'OTHER WOOD'
;
  value $woody
  ' '   = 'Missing'
  other = 'Specified'
;
 run;

title First CLASS Reduction for WOOD
Variable;
PROC FREQ data=scotch.scotch;
  tables wood/missing;
  format wood $wood.;
run;
```

```
title Second CLASS Reduction for WOOD Variable;
PROC FREQ data=scotch.scotch;
  tables  wood/missing;
  format wood $woody.;
run;

Title The Great Tabulate Output;
proc tabulate data=scotch.scotch noseps missing
                              formchar='          ';
  var age alcohol rating;
  table age alcohol rating
          ,
          N
          NMISS
         (min p5 p25 p50 p75 p95 max)*f=3. mean std
        /rts=20 row=float;
run;
TITLE;
```

For information of PROC FORMAT see (Bilenas and Tahilliani, 2019).

Some Exploratory Analysis (PLOTS)

BOX-WHISKER PLOTS or "BOX and Whisker Plots".

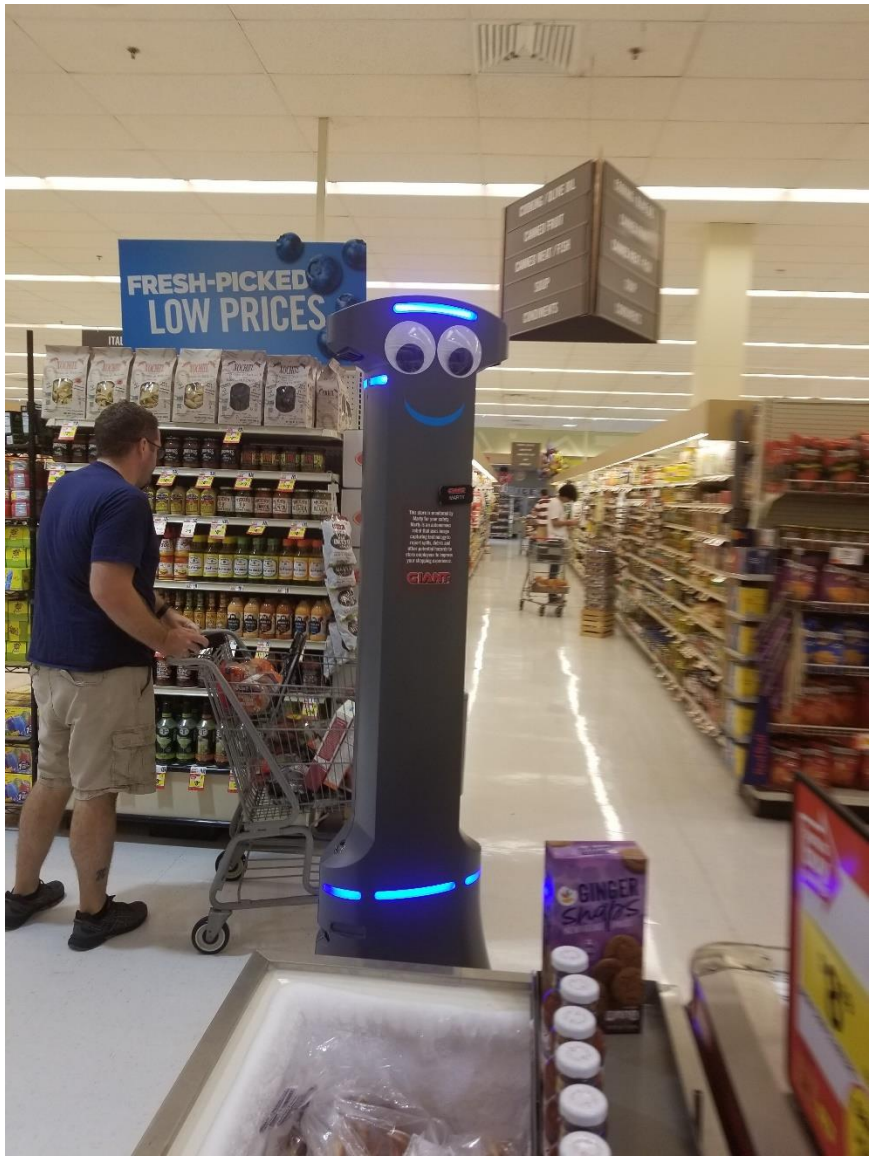
- **Component of EDA (Exploratory Data Analysis). A precursor to "Data Mining" developed at Bell Labs using S code, developed at Bell Labs.**
  - **John W. Tukey (1977) "Exploratory Data Analysis" Addison-Wesley Publishing.**
  - **In 1980, the first version of S was distributed outside of Bell Labs.**
  - **S-Plus and R are similar but may not be fully compatible with S.**
  - **S-Plus was bought by Tipco.com and is a commercial product called "Spotfire S+".**
- **https://en.wikipedia.org/wiki/S_(programming_language)**

EDA is Newer than Neural Networks.

- EDA developed in the 1970's.

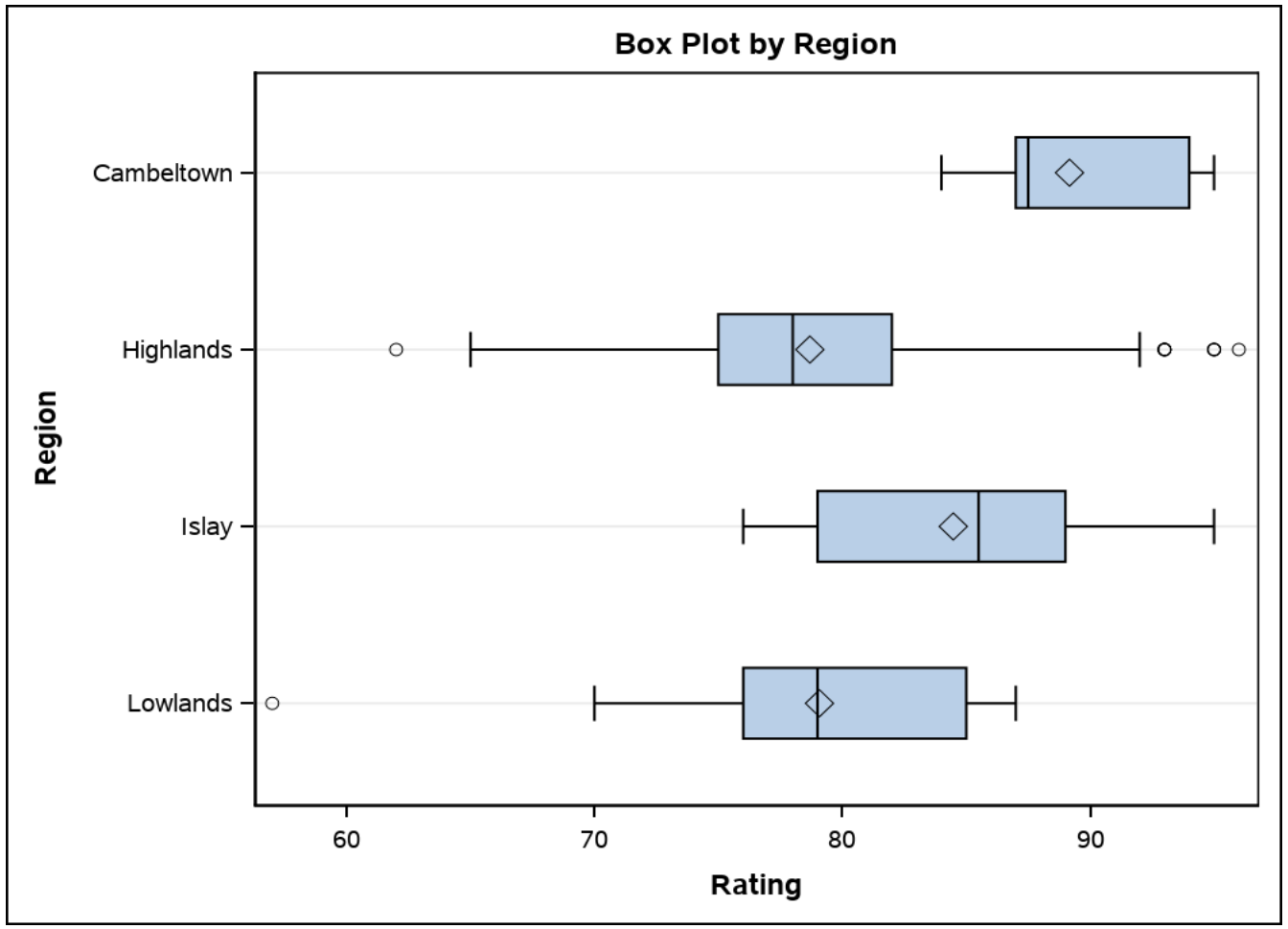
- Neural Networks first proposed in 1944. Interesting reference of the history and development of neural networks:
  - https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414
  - "The first trainable neural network, the Perceptron, was demonstrated by the Cornell University psychologist Frank Rosenblatt in 1957."

- "Terminator: Dark Fate" in theaters November 1st, 2019



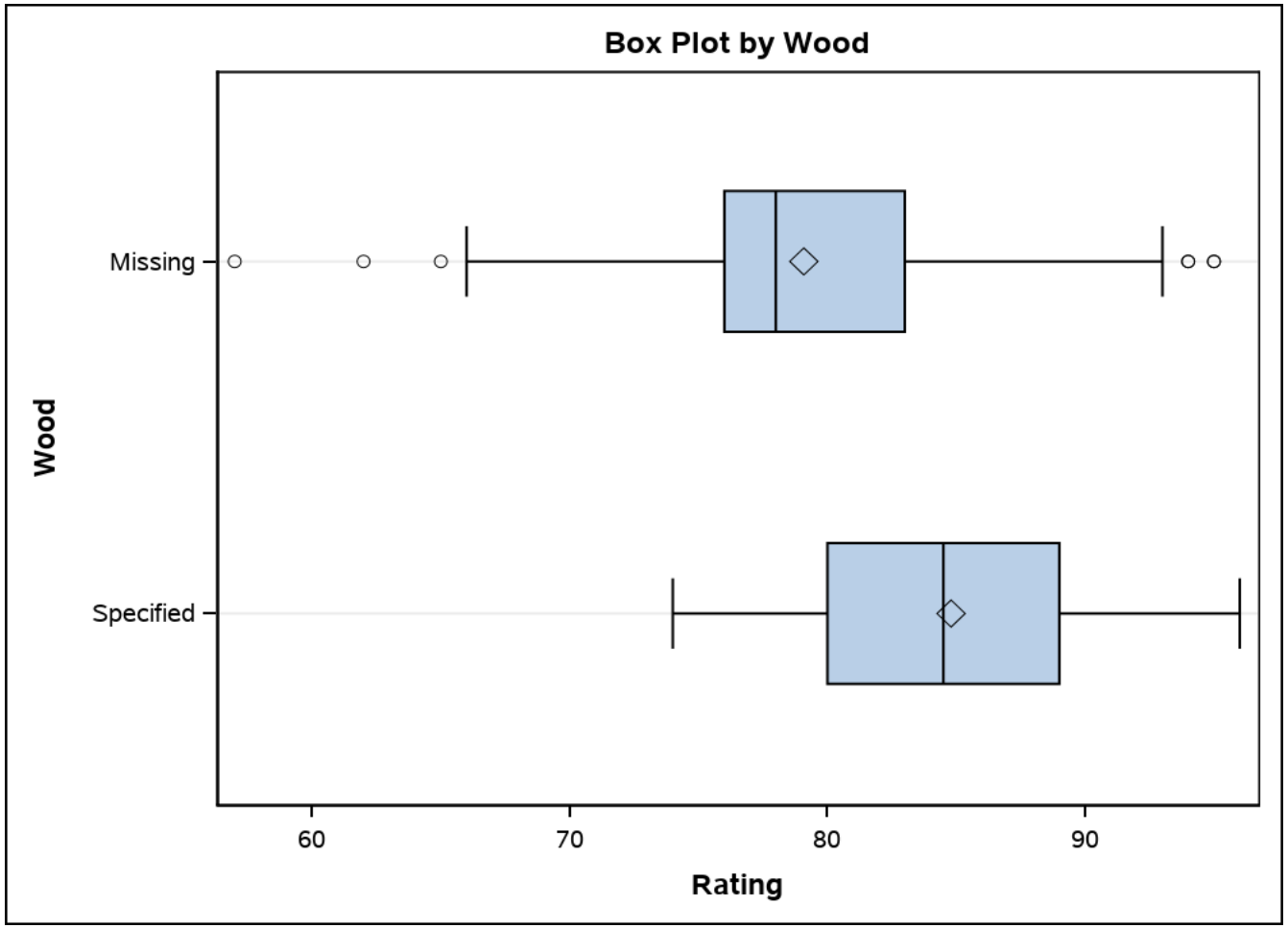
32

Some Exploratory Analysis (PLOTS)



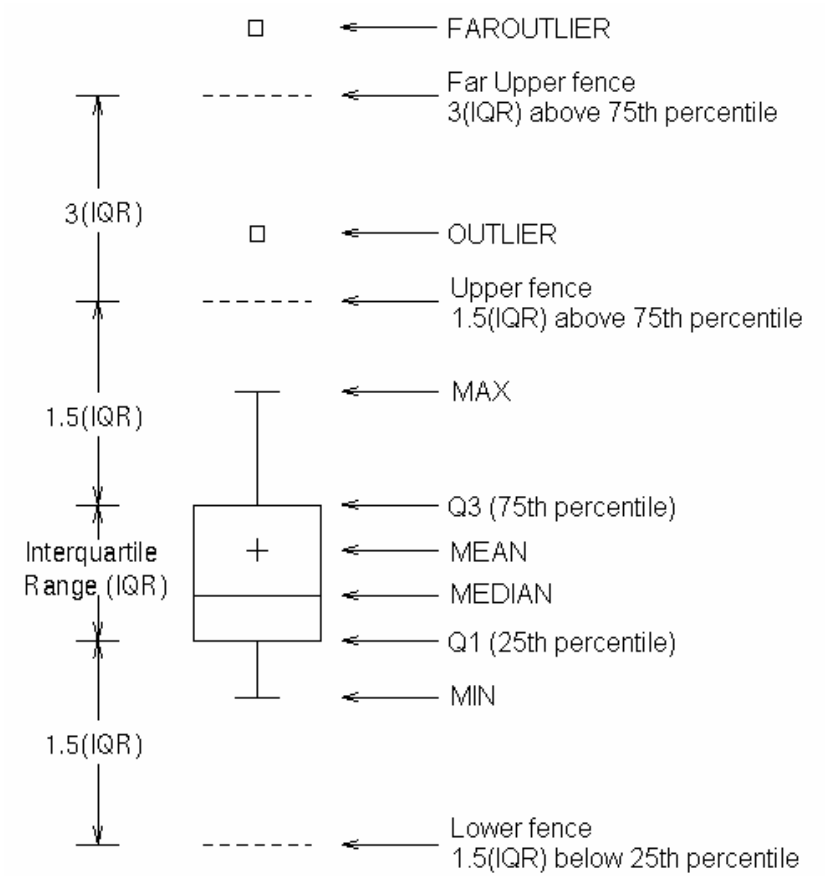

```
Title Some Box-Whisker Plots;
proc sgplot data=scotch.scotch;
  hbox rating / category=region
                missing;
  Yaxis grid label="Region";
  Xaxis label="Rating";
  title Box Plot by Region;
run;
```

Some Exploratory Analysis (PLOTS)



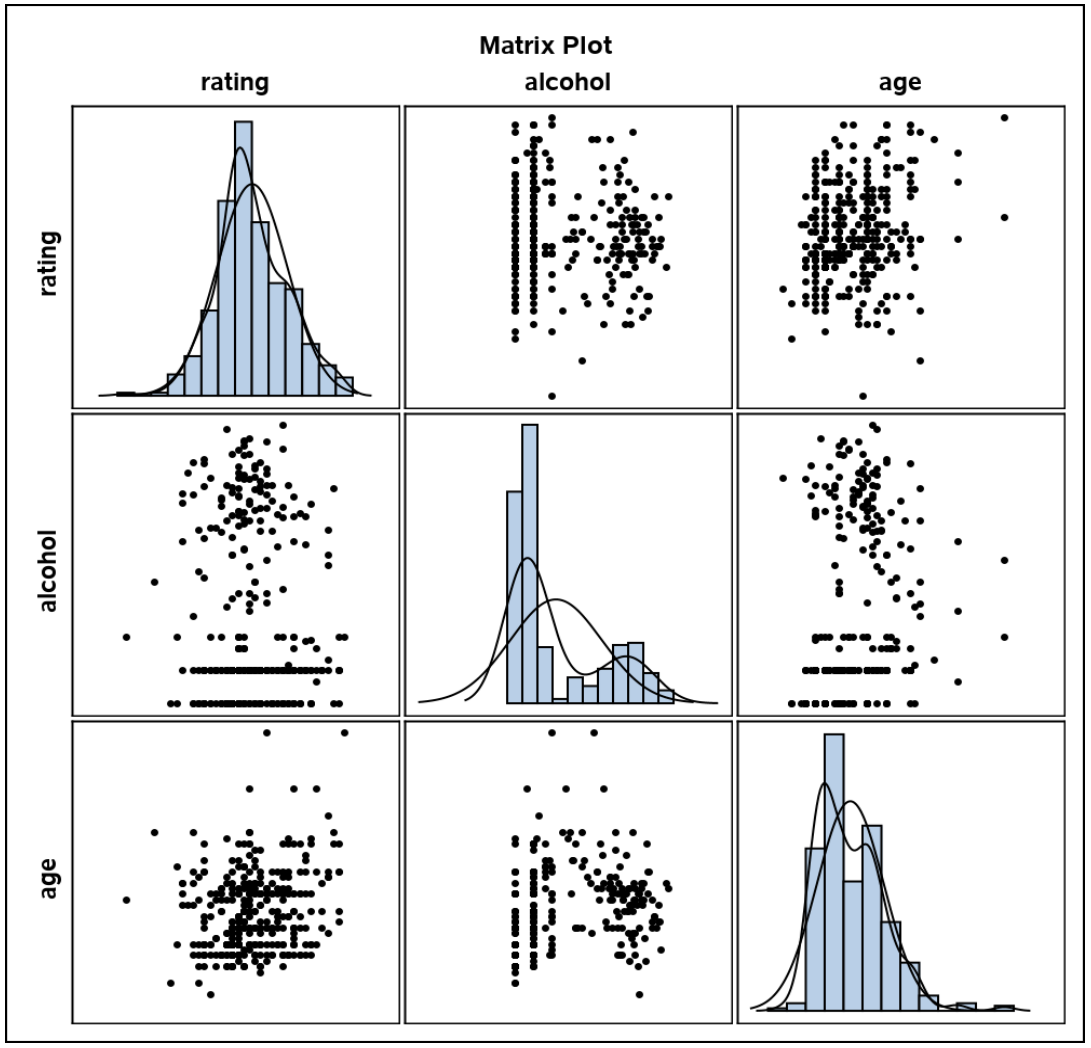

```
proc sgplot data=scotch.scotch;
  format wood $woody.;
  hbox rating / category=wood missing;
  Yaxis grid label="Wood";
  Xaxis label="Rating";
  title Box Plot by Wood;
run;
TITLE;
```

Some Exploratory Analysis (PLOTS)



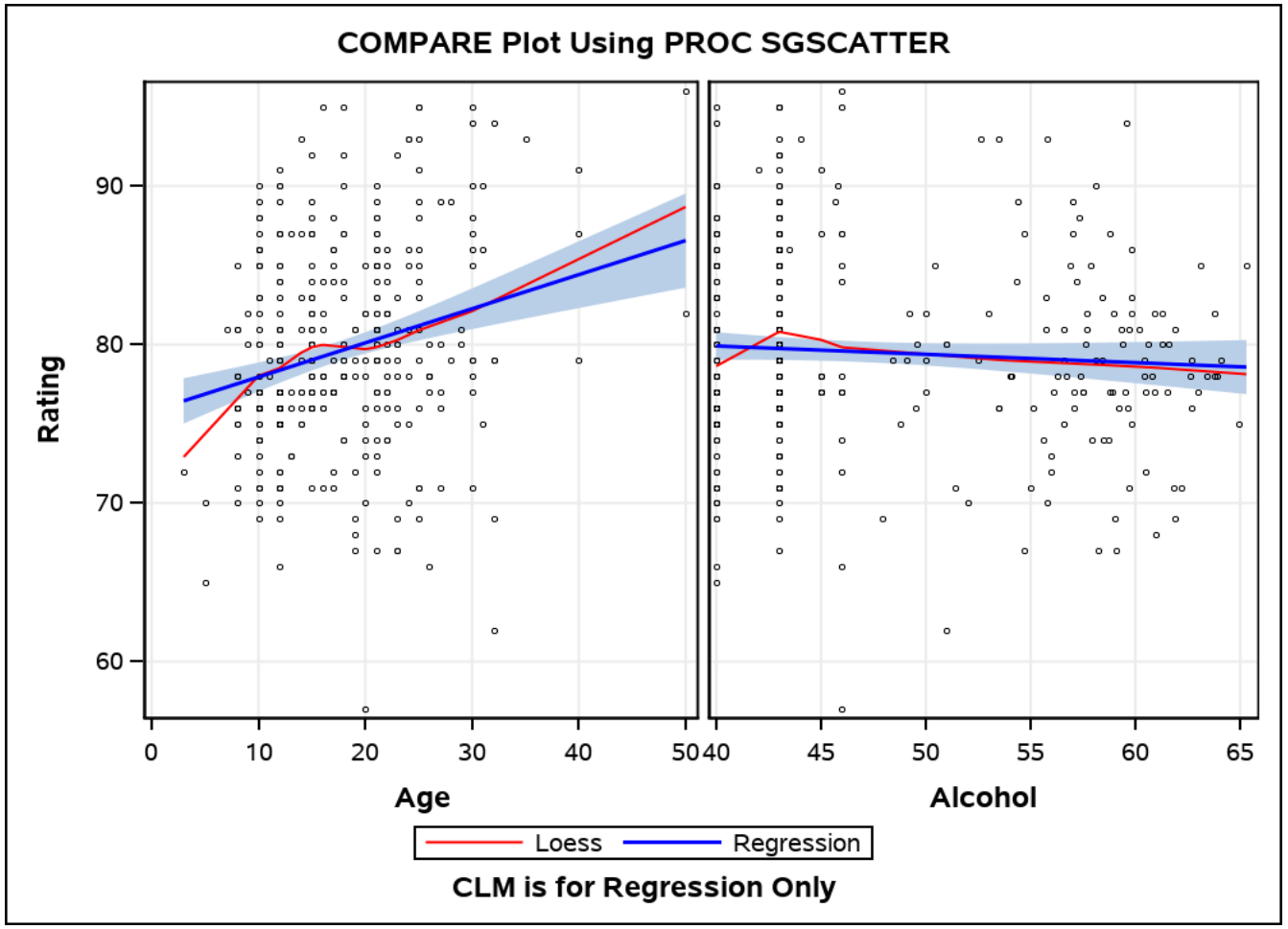
http://support.sas.com/documentation/cdl/en/grstatproc/62603/HTML/default/viewer.htm#vbox-stmt.htm

Some Exploratory Analysis (PLOTS)




```
Title MATRIX Plot Using PROC SGSCATTER;
proc sgscatter data=scotch.scotch;
        matrix Rating Alcohol Age
        / markerattrs=(symbol=circlefilled
                                /*size=2*/)
          diagonal=(histogram normal kernel);
run;
TITLE;
```

Some Exploratory Analysis (PLOTS)




```
Title COMPARE Plot Using PROC SGSCATTER;
proc sgscatter data=scotch.scotch;
  COMPARE y=Rating  x=(Age Alcohol)
        / markerattrs=(size=4 color=black) GRID
          LOESS=(smooth=0.5
                      lineattrs=(color=red thickness=.5))
          reg=(CLM lineattrs=(color=blue));
  label rating  = 'Rating'
          age     = 'Age'
          alcohol = 'Alcohol';
  footnote CLM is for Regression Only;
run;
TITLE;
FOOTNOTE;
```

Info on LOESS: see Bilenas & Herat (2016)

Collinearity tests

This section has been removed and maybe added to a future presentation.
For more information on Collinearity testing see:

- **Schreiber-Gregory, D. (2017)**
- **Belsley, Kuh, and Welsch R. (1980, 2004).**
- **Belsley (1991)**

Final Model using PROC GENMOD

Model Information	
Data Set	WORK.CENTERED
Distribution	Normal
Link Function	Identity
Dependent Variable	rating rating

Number of Observations Read	366
Number of Observations Used	366

Class Level Information			
Class	Value	Design Variables	
wood	Missing	0	
	Specified	1	
region	OTHER	0	0
	Islay	1	0
	Cambeltown	0	1

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	357	10378.5273	29.0715
Scaled Deviance	357	366.0000	1.0252
Pearson Chi-Square	357	10378.5273	29.0715
Scaled Pearson X2	357	366.0000	1.0252
Log Likelihood		-1131.4411	
Full Log Likelihood		-1131.4411	
AIC (smaller is better)		2282.8821	
AICC (smaller is better)		2283.5018	
BIC (smaller is better)		2321.9084	

Lowland region collapsed with Highland

- **AIC, AICC, BIC, and SBC are Information Criteria statistics that are better than R-Square or Adjusted R-squares.**
- **Not scaled to be from 0 to 1.**
- **The smaller the metric the better the model.**
- **Is dependent on the full data. You can't compare Information Criteria across development data sets.**
- **AIC and SBC are functions of number of observations, SSE (Sum of square Errors) and the number of independent variables which includes the intercept.**
- **See (Beal, 2007).**

Final Model USING PROC GENMOD

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	78.6633	0.3301	78.0162	79.3103	56776.7	<.0001
alc_center		1	-0.0912	0.0386	-0.1668	-0.0156	5.58	0.0181
age_center		1	0.2306	0.0411	0.1501	0.3110	31.51	<.0001
alc_center*age_center		1	-0.0235	0.0064	-0.0360	-0.0110	13.62	0.0002
wood	Specified	1	5.4235	1.0312	3.4024	7.4447	27.66	<.0001
region	Islay	1	5.3425	0.8892	3.5997	7.0853	36.10	<.0001
region	Cambeltown	1	10.1836	2.2111	5.8499	14.5173	21.21	<.0001
Scale		1	5.3421	0.1974	4.9688	5.7435		

Note: The scale parameter was estimated by maximum likelihood.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
alc_center	1	5.54	0.0186
age_center	1	30.23	<.0001
alc_center*age_center	1	13.37	0.0003
wood	1	26.66	<.0001
region	2	51.69	<.0001

Mean Alcohol=46.81863388 Mean Age= 17.426229508

Note SAS will provide under full rank model that Type3 Analysis is replaced with "**LR Statistics For Joint Tests** ":

Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests. The joint test for an effect is a test that all the parameters associated with that effect are zero. Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

Code Model USING PROC GENMOD STATS 15.10

```
proc format;
  value $woody
  ' '   = 'Missing'
  other = 'Specified'
;
  value $region
        'Cambeltown' = 'Cambeltown'
        'Islay'      = 'Islay'
        other        = 'OTHER'
;
run;

data _null_;
  set scotch.scotch end=eof;
  alc_mean ++ alcohol;
  age_mean ++ age;
  if eof then do;
    alc_center = alc_mean/_n_;
    age_center = age_mean/_n_;
    call symputx('alc_center',alc_center);
    call symputx('age_center',age_center);
  end;
run;
```

```
data centered;
  set for_reg1;
  alc_center=alcohol-&alc_center.;
  age_center=age-&age_center.;
  alc_age = alc_center*age_center;
run;

proc genmod data=centered NAMELEN=65;
  class wood region/order=freq param=ref ref=first missing;
  format wood $woody. Region $region.;
  model rating =alc_center|age_center@2 WOOD  REGION
              /dist=normal link=identity  type3;
  footnote Mean Alcohol=&alc_center. Mean Age= &age_center.;
run;
title; footnote;
```

Trivia Question: When Should You Run A Stepwise Regression?

a) Never.

b) Always.

c) When you want to take a long lunch break or vacation since the results may take a long time to produce, especially if running a stepwise Logistic Model.

Trivia Answers: When Should You Run A Stepwise Regression?

a) Never. *(5 points)*

b) Always. *(1 points)*

c) When you want to take a long lunch break or vacation since the results may take a long time to produce, especially if running a stepwise Logistic Model. *(2 points)*

Trivia: David Cassell's Answer from SAS-L:

http://listserv.uga.edu/cgi-bin/wa?A2=ind0610D&L=sas-l&P=R9019

Mon, 23 Oct 2006 13:52:28 -0700

The best time to use stepwise selection is when felons have had your family kidnapped, and are forcing you to do this at gunpoint. Oh wait, that's a Harrison Ford movie.

- Really, if you want to use stepwise selection, you have to be aware that it does not do what people want it to (i.e. magically come up with a 'best' set of predictor variables), and you need to go back and check the regression diagnostics for ***every* *single* intermediate stage** to make sure that it did not go drastically off-track because of one or more of the following:
  - Outliers
  - leverage points
  - non-normality of residuals
  - Heteroskedasticity
  - non-linearities
  - data contamination
  - mixtures of error distributions
  - multi-collinearity
  - Autocorrelation
  - suppressor variables . . . .
- The basic issue is that people think that stepwise selection will find a 'best' model with the right number of regressors. But it will not. The formulas for stepwise selection do not actually work. There is no checking for anything that can go wrong. So, even if your data are *perfect*, 100% multivariate normal errors with no outliers and no problems anywhere, you *still* cannot depend on stepwise selection to get you where you want to go.
- Frustrating, eh?

Also see; Flom, P.L. and Cassell, D.L. (2007) "Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use" NESUG 2007

Using SURVEY SELECT for VALIDATIONS

- **Bootstrap.**
- **K-Fold or Cross Validations.**

Trivia Question

- **We want to run a Bootstrap Analysis. Run the regression models 1000 times based on each of the Bootstrap samples generated with a 100% sample rate.**
- **What sampling METHOD should be used?**
    - **a) SRS: simple random sampling without replacement.**
    - **b) URS: unrestricted random sampling with replacement**
    - **c) BALBOOTSTRAP which uses unrestricted sampling with replacement**
    - **d) B or C**

Trivia Answers

a) SRS: simple random sampling without replacement. 0 points.

a) URS: unrestricted random sampling with replacement. 3 points.

b) BALBOOTSTRAP which uses unrestricted sampling with replacement. 3 points

c) B or C. 6 points

Two Bootstrap Runs: 1,000 samples

Bootstrap using METHOD=URS

Parameter	Level1	mean	ci0_5	ci2_5	ci5	ci50	ci95	ci97_5	ci99_5	p
Intercept		78.665259737	77.849470502	78.067109961	78.134356635	78.669890078	79.216326498	79.300230374	79.459692093	0.01
age_center		0.2261124192	0.1159523155	0.1370426087	0.1530653508	0.2267892418	0.2933008319	0.3041252157	0.3248596557	0.01
alc_center		-0.091536587	-0.193853371	-0.161423583	-0.149477507	-0.091124995	-0.031890424	-0.02144191	-0.002662988	0.01
alc_center*age_center		-0.024533272	-0.04160555	-0.037215894	-0.035065978	-0.024523164	-0.014472774	-0.012865683	-0.009923555	0.01
region	Cambeltown	10.279906	5.8922469632	6.7476421265	7.090263911	10.196686241	13.525964887	14.52140256	16.547855522	0.01
region	Islay	5.361493671	3.34550367	3.7176183256	3.9144326687	5.3345367919	6.8236614259	6.9841699254	7.8984878021	0.01
wood	Specified	5.3814714315	2.4885074426	3.3291165555	3.6180395488	5.3732937245	7.1421050757	7.4280183798	8.1547120544	0.01

Bootstrap using METHOD=BALBOOT

Parameter	Level1	mean	ci0_5	ci2_5	ci5	ci50	ci95	ci97_5	ci99_5	p
Intercept		78.667656312	77.811530829	77.990520662	78.075377429	78.667677107	79.211197439	79.323189059	79.497583901	0.01
age_center		0.2268830143	0.1141333078	0.1430833819	0.1540551823	0.2286104225	0.2957422473	0.308348406	0.3329824788	0.01
alc_center		-0.091124832	-0.186237288	-0.161712325	-0.151336949	-0.091034435	-0.033275248	-0.019468989	-0.009163644	0.01
alc_center*age_center		-0.024137148	-0.041726578	-0.036956392	-0.034551026	-0.024124775	-0.013640066	-0.011727704	-0.007462545	0.01
region	Cambeltown	10.158151689	5.9651391479	6.6399071124	7.0185556212	10.077069783	13.477523161	13.953528632	16.168679556	0.01
region	Islay	5.3636560515	3.2142196033	3.64677703	3.9040815458	5.3843262909	6.8274156657	7.0383527604	7.3443657789	0.01
wood	Specified	5.4346084605	2.5213100913	3.3484607947	3.5982515606	5.4221176288	7.2470003915	7.5054626117	8.1836382971	0.01

Bootstrap Runs: 1,000 samples Code

```
%macro break;
ods EXCEL options(sheet_interval="output");
ods exclude all;
data _null;
  declare odsout obj();
run;
ods select all;
%mend break;

%let outp=~;
Libname scotch "&outp";

%let level = level1;
%*let level = level1 level2; /* 2 way interactions */
/* leave blank if no class variables  */

proc format;
  value $wood
  ' '      = 'Missing'
  'Sherry' = 'SHERRY'
   other   = 'OTHER WOOD'
;
  value $woody
  ' '   = 'Missing'
  other = 'Specified'
;
  value $region
        'Cambeltown' = 'Cambeltown'
        'Islay'      = 'Islay'
        other        = 'OTHER'

 run;
```

```
data _null_;
  set scotch.scotch end=eof;
  alc_mean ++ alcohol;
  age_mean ++ age;
  if eof then do;
    alc_center = alc_mean/_n_;
    age_center = age_mean/_n_;
    call symputx('alc_center',alc_center);
    call symputx('age_center',age_center);
    file print;
    put alc_center=
       /age_center=;
  end;
run;

data center;
  set scotch.scotch;
  alc_center=alcohol-&alc_center.;
  age_center=age-&age_center.;
run;
```

This code only produces 1 bootstrap report using the URS METHOD using some code from (Cassell , 2007)

Bootstrap Runs: 1,000 samples Code

```
SASFILE center LOAD;
proc surveyselect data=center
                  out=outdata
                  seed=20191022
                  rep=1000
                  method=URS
                  samprate=1 /* 100% rate */
                  outhits;
run;
SASFILE center CLOSE;

proc printto log='/~/SURVEY_PROCS/bs.log';

ods graphics off;
ods exclude all;
ods noresults;
```

```
proc genmod data=outdata namelen=65;
  ods output ParameterEstimates=bout;
  by replicate;
  class wood region/order=freq param=ref
                    ref=first missing;
  format wood $woody. region $region.;
  model rating =
       alc_center|age_center@2 WOOD  REGION;
run;
ods output close;

ods graphics on;
ods exclude none;
ods results;

proc printto log=log;
```

```
/* FOR BALBOOTSTAP Results*/
proc surveyselect data=center
                  out=outdata2
                  seed=20191022
                  rep=1000
                  method=BALBOOT
                  outhits;
run;
```

Bootstrap Runs: 1,000 samples Code

```
proc sort data=bout force noequals tagsort;
  by parameter &level.;
run;

proc univariate data=bout noprint;
  by parameter &level.;
  var estimate;
  output out=final pctlpts=0.5, 2.5, 5, 50, 95,
97.5, 99.5 pctlpre=ci
         mean=mean;
run;


data ci;
  set final;
  if sign( ci0_5) = sign( ci99_5) then p=0.01;
  else if sign(ci2_5) = sign(ci97_5) then p=0.05;
  else if sign(ci5)   = sign(ci95) then p=0.10;
  else p=.;
run;
```

```
ods EXCEL file="&outp./validations.4.xlsx"
        style=SASWEB
          OPTIONS (fittopage = 'yes'
                   frozen_headers='no'
                   autofilter='none'
                   embedded_titles = 'yes'
                   embedded_footnotes = 'yes'
                   zoom = '100'
                   orientation='Landscape'
                   Pages_FitHeight = '100'
                   center_horizontal = 'no'
                   center_vertical = 'no'
                );

ods EXCEL options(sheet_interval="none"
                   sheet_name="Bootstrap"
                  );
```

Bootstrap Runs: 1,000 samples Code

```
proc format;
  value p  .01 = 'Lime'
           .05 = 'Green'
           .10 = 'Yellow'
        .  = 'Red'
;
run;

proc print data=ci (where=(parameter ne 'Scale')) noobs;
  format parameter $65. _numeric_ best12.;
  var Parameter  &level.  mean
                      ci0_5 ci2_5 ci5 ci50 ci95 ci97_5 ci99_5;
  var p / style={background=p.};
  title "Bootstrap ";
run;
%break;
```

K-FOLD Validations

- **A hold-out sample is not really a K-Fold or Cross Validation.**
- **For K-FOLD we generate K samples of the data and then each K sample uses (K-1)/K to build the model and then test out the remaining 1/K of the sample with the model to determine if the fit is ok.**
- **The example we will take is K=3, resulting in 3 validations. Model is built on 2/3 of the sample. The K "non-selected" observations will have the Dependent Variable set to missing but still will be included in the GENMOD model. They will not be included in the model build but will get a prediction of the DV from running in the regression procedure.**
- **We will then review results on the K samples.**
- **Most of the code is from (Cassell, 2007)**

K-FOLD Validations CODE

```
ods EXCEL options(sheet_interval="none"
          sheet_name="K-FOLD"
         );

/* K-FOLD Data */
%let K=3;
%let rate=%sysevalf((&K-1)/&K);
%let y=rating;

/* generate the cross-validation sample */
proc surveyselect data=center out=xv seed=1874
        samprate=&RATE outall rep=3;
     * outall - Retains all records. Those selected get SELECTED=1.  Not selected get SELECTED=0 */
 run;

data xv;
 set xv;
 if selected then new_y=&y.;  /* Sets K sample DV to missing */
 run;
/* get predicted values for the missing new_y in each replicate */
proc genmod data=xv namelen=65;
  class wood region/order=freq param=ref ref=first missing;
  format wood $woody. region $region.;
  model new_y =alc_center|age_center@2 WOOD  REGION
            /WALD type3;
  by replicate;
  output out=out1(where=(new_y=.)) p=yhat;
run;
/* end K-FOLD set-up */
```

K-FOLD Validations CODE

```
/* K-Fold  Results */
/* summarize the results of the cross-validations */
data out2;
 set out1;
 d= &y. - yhat;
 absd=abs(d);
 run;

proc tabulate data=out2 missing noseps;
  class replicate;
  keylabel mean=' ' std= ' ' var=' ';
  var d absd;
  table replicate
        all
        ,
        d = 'Deviance'*var*f=best12.
        d = STD Deviance'*std*f=best12.
        absd = 'MAE'*mean*f=best12.
        /rts=12 row=float misstext='';
title K-FOLD Cross Validation;
run;

proc sgpanel data=out2;
  panelby replicate;
  histogram d;
  density d/type=normal;
  density d/type=kernel;
  label d='Residual';
 run;
Title;
%break;
```

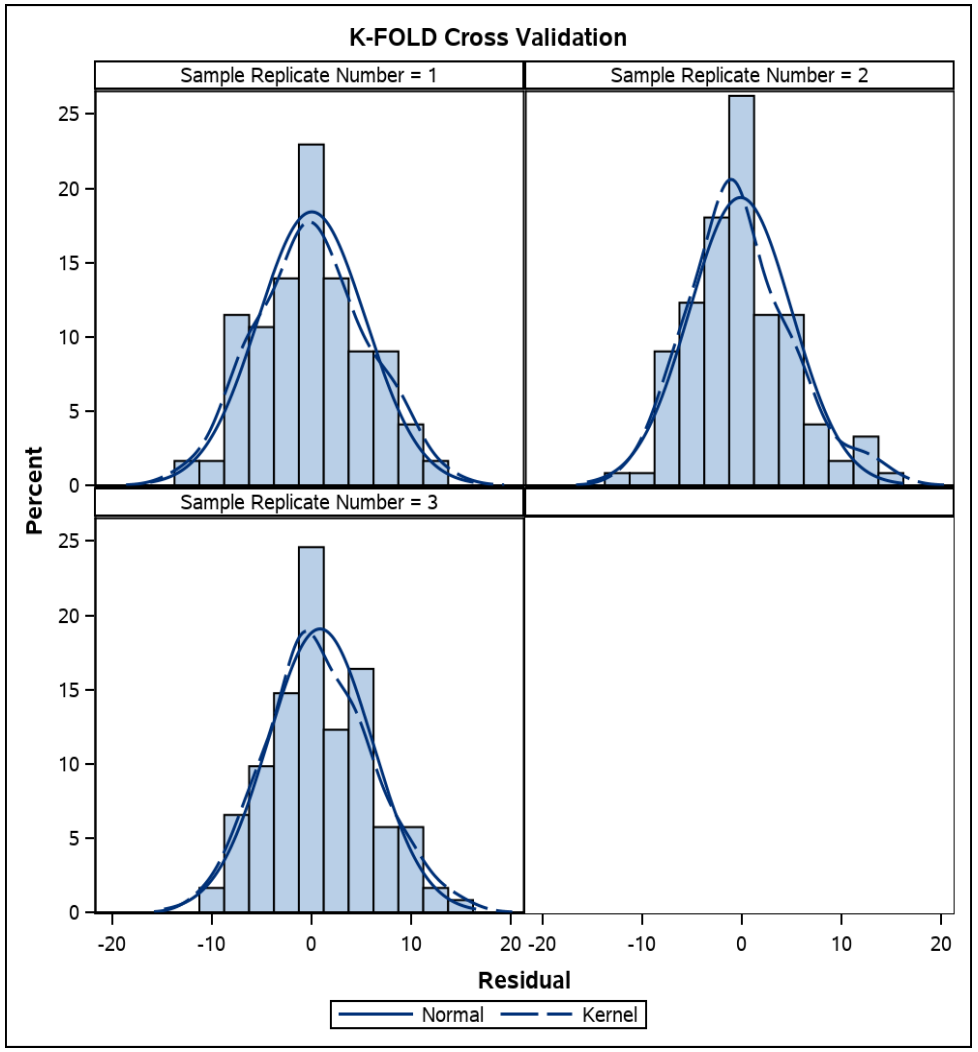
K-FOLD Validations Results

Selection Method	Simple Random Sampling
Input Data Set	CENTER
Random Number Seed	1874
Sampling Rate	0.66666667
Sample Size	244
Selection Probability	0.666667
Sampling Weight	0
Number of Replicates	3
Total Sample Size	732
Output Data Set	XV

K-FOLD Validations Results

K-FOLD Cross Validation

Sample Replicate Number	Deviance	STD Deviance	MAE
1	29.335079121	5.4161867694	4.2757814335
2	26.480367441	5.1459078345	3.9732420487
3	27.311238839	5.2260155797	4.1471576948
All	27.743829257	5.2672411428	4.1320603923



SURVEYREG vs REG

- Sampling records and then running a regressions should be done in PROC SURVEYREG or PROC SURVEYLOGISTIC.
- Using a weight option in PROC REG or LOGISTIC will not give you the correct standard errors and as a result, the p-values will be incorrect.
- Lets see results for a simpler model.

SURVEYREG vs REG CODE

```
proc sort data=scotch.scotch out=region
        noequals tagsort force;
  by region;
run;

proc surveyselect data=region
                  method=srs
                  rate=(1 0.1 0.4  0.5)
                  out=sample2
                  seed=20191022;
  strata region;
run;


title PROC REG;
proc reg data=sample2 plots=none;
  model rating = age;
  weight SamplingWeight;
  title REG for SAMPLED DATA;
run;
title;
```

```
data strat_totals;
  input region $ _TOTAL_;
  datalines;
Cambeltown     6
Highlands     292
Islay          42
Lowlands       26
;;
run;


proc surveyreg data=sample2
            total=strat_totals
            plots=none;
  strata region / list;
  model rating = age;
  weight SamplingWeight;
  title PROC SURVEYREG with total Ns;
run;
title;
```

SURVEYREG vs REG CODE

```
proc surveyreg data=sample2
              plots=none;
  strata region / list;
  model rating = age;
  weight SamplingWeight;
  title PROC SURVEYREG without total Ns;
run;
title;
```

SURVEYREG vs REG Output

REG for SAMPLED DATA

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	74.92565	1.68591	44.44	<.0001
age	age	1	0.25474	0.08093	3.15	0.0025

PROC SURVEYREG with total Ns

Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	74.9256546	1.91167650	39.19	<.0001
age	0.2547427	0.10430972	2.44	0.0175

PROC SURVEYREG without total Ns

Estimated Regression Coefficients				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	74.9256546	2.04398957	36.66	<.0001
age	0.2547427	0.11140194	2.29	0.0256

References

- **Beal, D.J. (2007). "Information Criteria Methods in SAS® for Multiple Linear Regression Models", SESUG 2007.** https://analytics.ncsu.edu/sesug/2007/SA05.pdf
- **Belsley, D., Kuh, E., and Welsch, R. (2004). "Regression Diagnostics, Identifying Influential data and Sources of Collinearity", Wiley (reprint of 1980 edition).**
- **Belsley, D. (1991). "A Guide to Using the Collinearity Diagnostics" Computer Science in Economics and Management 4: 33-50.**
- **Bilenas, J. (2009). "Using the New SURVEY Procedures from a Modeling Perspective" NESUG 2009**. https://www.lexjansen.com/nesug/nesug09/sa/SA12.pdf
- **Bilenas, J. and Herat, N. (2016). "Using Regression Splines in SAS® STAT Procedures" SESUG 2016** https://analytics.ncsu.edu/sesug/2016/BF-140_Final_PDF.pdf
- **Bilenas, J. and Tahiliani, K. (2019). "The Power of the PROC FORMAT" PharmaSUG 2019** https://www.lexjansen.com/pharmasug/2019/BP/PharmaSUG-2019-BP-057.pdf
- **Bilenas, J. and Tahiliani, K. (2016). "Making Sense of PROC TABULATE" SESUG 2016**. https://analytics.ncsu.edu/sesug/2016/HOW-138_Final_PDF.pdf
- **Cassell, David. L (2007). "Don't Be Loopy: Re-Sampling and Simulation the SAS® Way". SAS GLOBAL FORUM. http://www2.sas.com/proceedings/forum2007/183-2007.pdf**
- **Cassell, D. L. (2006) "Wait , Don't Tell Me… You're Using the Wrong Proc!" SUGI31. Paper 193-31.**
- **Effron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 1:54-77.**
- **Jackson, M. (1999), Michael Jackson's Complete Guide to Single Malt Scotch, 4th Edition. Running Press.**
- **Flom, P.L. and Cassell, D.L. (2007) "Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use" NESUG 2007**
- **WICKLIN, RICK (2013). "SIX REASONS YOU SHOULD STOP USING THE RANUNI FUNCTION TO GENERATE RANDOM NUMBERS." AVAILABLE AT https://blogs.sas.com/content/iml/2013/07/10/stop-using-ranuni.html**
- **Schreiber-Gregory, D. (2017) "Multicollinearity: What Is It, Why Should We Care, and How Can It Be Controlled?" SESUG 2017.** https://analytics.ncsu.edu/sesug/2017/SESUG2017_Paper-160_Final_PDF.pdf
- **Tukey, J.W. (1977("Exploratory Data Analysis 1st Edition." Addison-Wesley Publishing Company.**

References

Michael Jackson

March 27, 1942 – August 30, 2007

Interesting Websites to Explore Whiskey or Whisky

- http://www.manatawnystillworks.com/
  - Distillery and bar in Pottstown.
  - 320 Circle of Progress Drive
    Suite 104
    Pottstown, PA 19464
  - Saturday tours @ 1:00 pm and 5:00 pm
  - Sunday tours @ 2:00 pm

- https://thewhiskeyjug.com/japanese-whiskey/ohishi-whisky-sherry-cask-review/
  - Interesting Japanese Rice Whisky aged in Sherry Casks

Disclaimers

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names used in this presentation are trademarks of their respective companies.

The contents of this paper are the work of the author and do not necessarily represent the opinions, recommendations, or practices of any company that I have worked for or are currently working for.

No warranty for any code in this presentation. Use at your own risk.

All code was generated and tested on SAS Studio SAS® OnDemand for Academics

I did not receive any advertising proceeds for Whisky recommendations made in this presentation.